Smerdiagina, Anna

Article

Lost in transcription: Experimental findings on ethnic and age biases in AI systems

Junior Management Science (JUMS)

Provided in Cooperation with:
Junior Management Science e. V.

Suggested Citation: Smerdiagina, Anna (2024) : Lost in transcription: Experimental findings on ethnic and age biases in AI systems, Junior Management Science (JUMS), ISSN 2942-1861, Junior Management Science e. V., Planegg, Vol. 9, Iss. 3, pp. 1591-1608, https://doi.org/10.5282/jums/v9i3pp1591-1608

This Version is available at:
https://hdl.handle.net/10419/305308

Junior Management Science 9(3) (2024) 1591-1608

Junior Management Science

www.jums.academy
ISSN: 2942-1861

Lost in Transcription: Experimental Findings on Ethnic and Age Biases in AI Systems

Anna Smerdiagina

Technical University of Munich

Abstract

Artificial intelligence (AI) has revolutionized industries and improved our lives in various ways. However, AI systems' potential to amplify existing biases in society has become a major concern. This thesis explores the concept of bias in AI and how it can lead to discrimination, focusing specifically on the performance of Automatic Speech Recognition (ASR) systems in relation to the ethnicity (accent) of participants. The study collected 187 recordings from proficient English speakers of 55 ethnic groups.These recordings were transcribed via ASR systems and evaluated by the word error rate (WER) metric. The ASR systems selected for the study were Gboard (Android) by Google, Apple keyboard (iOS), and Whisper by Open AI. The study results show that ethnicity significantly impacts the performance of ASR systems, with some ethnic groups experiencing substantially higher error rates than others. The study provides evidence that ASR systems may not be equally accurate for all users. To address ethnic bias in AI systems, it is essential to take a multi-faceted approach involving technical and societal solutions. The findings highlight the importance of addressing bias in AI systems to ensure fairness, transparency, and equity for all users, regardless of ethnicity.

Keywords: automatic speech recognition; bias in AI; digital ageism; digital equity; ethnic bias

1. Introduction

Artificial intelligence (AI) is revolutionizing many industries and making our lives easier in various ways (Bostrom, 2014). However, bias in AI systems has become a major concern in recent years due to the potential of perpetuating and amplifying biases that are already present in society. Bias in AI can occur at various stages of the development process, including the selection of training data, the design of the algorithms, and the interpretation of the results (Mehrabi et al., 2021). One particular area of concern is the potential for AI systems to exhibit biases based on ethnicity and age, which can result in discrimination against certain groups of people (Barocas & Selbst, 2016).

The impact of bias in AI can be significant, as these systems are increasingly being used in a variety of contexts, including hiring, lending, and criminal justice (Kleinberg et al., 2018). For example, a biased AI system that is used in the hiring process may unfairly reject job candidates from certain ethnicities or age groups. Similarly, a biased AI system that is used in the criminal justice system may disproportionately affect certain groups of people, leading to further inequalities and injustices.

In this thesis, we will explore the concept of bias in AI and how it can lead to discrimination. We will review the literature on the ways in which AI systems can exhibit biases based on ethnicity and age. We will consider the impact that these biases can have on individuals and society as a whole, and explore potential ways to mitigate them. Further, we will conduct an experiment to investigate the extent to which these biases exist in specific AI systems.

The hypothesis of this thesis is that the performance of automatic speech recognition (ASR) systems may be influenced by the ethnicity (accent) of participants. We will test this

I sincerely thank Prof. Dr. Isabell Welpe, my academic guide, whose unwavering trust in my creative journey and agile supervision has been an invaluable compass. My heartfelt thanks to the 210 participants whose voices enriched the exploration of fairness in automatic speech recognition. Special appreciation goes to Edman Paes dos Anjos for his constant support and belief in me.

DOI: https://doi.org/10.5282/jums/v9i3pp1591-1608

hypothesis on three ASR systems: Gboard by Google, Apple keyboard by Apple Inc., and Whisper by OpenAI, and compare the results of native English speakers with the results of non-native English speakers from various ethnic groups.

The research questions of this thesis are as follows:

1. How does the bias in AI systems affect automatic speech recognition for different ethnic groups?
2. Does the ethnicity (accent) of a user affect the performance of automatic speech recognition systems?
3. What are the best practices to reduce bias in AI systems, specifically in the context of ethnicity and age?

Overall, this thesis aims at contributing to the understanding of biases in AI and providing insights on how they can be addressed in order to create more equitable AI systems. We hope to raise awareness of the potential consequences of bias in AI for vulnerable groups and provide recommendations for addressing biases in the development and deployment of AI systems.

The thesis is structured as follows: we will start by providing a theoretical background on bias in AI, including different types of biases and how they can occur. We will then examine the literature on ethnicity and age bias in AI systems. Further, we will discuss measures for reducing these biases in AI systems. The methodology chapter will describe the experiment we conducted to investigate the extent to which these biases exist in the specific AI system. Finally, the results and discussion chapter will present and analyze the results of the experiment and discuss their implications.

2. Definition of Bias

At any moment in time, there is a stream of 11 billion bits of information coming to us from every sense that we have. The human mind is only able to consciously process 40 bits (Zimmermann, 1986). This implies that most of our decisions are subconscious or unconscious. Given the overwhelming amount of information, the nervous system is only able to function through the use of cognitive shortcuts, also known as heuristics (Storage, 2021). However, these shortcuts can instigate discriminative behavior.

In this chapter, we will examine the types of biases that can manifest in AI systems, including cognitive biases, biases in machine learning, and biases in AI speech recognition. We will also explore the sequence of stereotypes, prejudice, and discrimination and how it leads to biased outcomes. The main focus of this chapter is to provide a comprehensive understanding of the various forms of biases that can manifest in AI systems and differentiate key definitions.

2.1. Cognitive Bias and Its Implications

It is essential to differentiate the key definitions to understand where and how biases arise. The concept of "cognitive misers," or the tendency to rely on mental shortcuts when making decisions due to limited cognitive resources, can contribute to biases (Fiske & Taylor, 1991). These mental shortcuts can lead to the automatic activation of stereotypes and biases, causing biased decisions and outcomes (Kleinberg, 2018).

Cognitive biases can manifest through the sequence of a stereotype, prejudice, and discrimination (Dasgupta & Asgari, 2004; Storage, 2021). Stereotypes are beliefs about the characteristics of a group of people, and these beliefs can be based on a variety of factors such as race, gender, sexual orientation, and religion (Dovidio, 2001). Stereotypes can be either positive or negative, and can lead to negative attitudes, or prejudice, towards certain groups (Dovidio, 2001; Storage, 2021). For instance, the stereotype that men are more capable and competent in the workplace than women can lead to prejudice towards women in the workplace, resulting in discriminatory behaviors such as not promoting women to leadership positions or paying them less than their male counterparts (Eagly & Karau, 2002).

In contrast with a stereotype, prejudice can only be negative (Storage, 2021). Prejudice, or negative attitudes towards a group of people, can then manifest in discriminatory behavior towards a certain group (Kleinberg, 2018). For instance, prejudice against individuals with disabilities might lead to discrimination, such as denying employment or education opportunities (Olkin & Pledger, 2003). The sequence of stereotypes, prejudice, and discrimination can aggravate existing societal biases and disparities (Barocas & Selbst, 2016).

One of the ways that biases can manifest is through explicit and implicit biases (Greenwald & Krieger, 2006). Explicit biases are conscious and intentional, and people may be aware of their own explicit biases (Greenwald et al., 2015). For instance, a study showed that Uber and Lyft drivers were canceling rides or extending wait times for African–American customers based on their names and faces upon the order, which is a direct and intentional form of discrimination (Ge et al., 2016). Additionally, the study found that women were taken on longer routes to extend the cost of the fare, also a direct indication of explicit bias. According to Lee (2018), explicit biases can and must be traced and mitigated further by law reinforcement.

Implicit biases, on the other hand, are unconscious and automatic (Greenwald et al., 2015). Implicit biases can be particularly insidious as they are not always recognized or acknowledged, yet they can still influence behavior (Nosek et al., 2002). This case is illustrated in an experiment that showed that using word embeddings in machine learning processes can lead to sexist results (Bolukbasi et al., 2016). For instance, in word analogy tests, "man" would be assigned to "computer programmer" while "woman" would be assigned to "homemaker." This bias toward women triggered the authors to propose a method that respects the embeddings for gender-specific words but de-biases embeddings for gender-neutral words.

In conclusion, cognitive biases in AI systems have the potential to exacerbate existing societal issues, causing biased outcomes. Understanding the different types of biases that can occur in machine learning is crucial in developing effec-

tive strategies for mitigating these biases. The next section will delve deeper into these specific types of biases in AI systems.

2.2. Types of Biases in Machine Learning

Machine learning, as a subfield of artificial intelligence, has become an integral part of a human routine, from food delivery to airport security procedures, affecting every individual in various ways (Guegan & Hassani, 2018; Guimaraes & Tofighi, 2018). However, one of the major challenges facing machine learning is the presence of biases in the data that is used to train these models, as well as flawed training and testing processes. These biases can lead to unfair and inaccurate outcomes, particularly for marginalized groups. In this section, we will explore the different types of biases that can occur in machine learning and the methods that can be used to address them.

2.2.1. Data bias

Data bias refers to the systematic errors or distortions that occur when the data used to train or evaluate machine learning models is unrepresentative or skewed in some way (Baeza-Yates, 2018). Data bias can be caused at any phase in a range of areas, from human reporting and selection bias to annotator bias (Hellström et al., 2020). The use of AI systems that are trained on biased data has the potential to amplify harmful stereotypes about certain ethnicities. For instance, an AI system trained on data that includes negative stereotypes about certain ethnicities may influence the way individuals are treated or perceived, escalating inequality.

2.2.2. Sampling bias

In the field of machine learning, sampling bias occurs when the sample of data used to train a machine learning model is not representative of the population it is intended to model (Mehrabi et al., 2021). If a model is trained on data that is predominantly from one gender or race, it may not accurately reflect the characteristics of the broader population and may lead to biased results. In 2018, Reuters reported that an AI system used to evaluate job applicants by Amazon's Human Resources department was biased to advise hiring male candidates, resulting in fewer female individuals being offered job opportunities (Dastin, 2018).

2.2.3. Selection bias

Selection bias occurs when the data used to train a model is selected in a non-random manner, resulting in a sample that is not representative of the population (Shah et al., 2020). This can occur when data is self-selected, such as in online surveys, or when data is selected based on certain criteria, such as data from only a particular geographic region (Baeza-Yates & Ribeiro-Neto, 2011).

2.2.4. Measurement bias

Measurement bias refers to errors or distortions in the way data is collected, recorded, or measured (Suresh & Guttag, 2019). For instance, if data is collected using a biased survey instrument or by a researcher with a preconceived notion about the outcome, the resulting data may be biased (Hajian et al., 2016).

2.2.5. Label bias

Label bias, or annotator bias (Hellström et al., 2020), refers to inconsistent labeling processes: when different annotators have mismatching styles that lead to misunderstanding and get reflected in the labels created. A common occurrence of label biases happens when differing labels get assigned to the same type of object by different annotators (for instance, grass vs. lawn, painting vs. picture) (Malisiewicz & Efros, 2008).

2.2.6. Confirmation bias

Confirmation bias is a type of cognitive bias that occurs when people seek out or interpret information in a way that confirms their preexisting hypotheses or opinions. In the context of machine learning, confirmation bias can occur when data is selected or analyzed in a way that confirms the researcher's expectations or hypotheses, leading to partial results (Carvalho et al., 2019). Some researchers recognize confirmation bias as a sub-type of a label bias (Srinivasan & Chander, 2021).

2.2.7. Negative Set bias

Negative set bias refers to the unreasonable emphasis on negative examples (examples that the model is attempting to classify as a particular class) in comparison to positive examples (examples that are not being classified as that particular class). As a result, datasets that only collect data on negative instances might be biased and disadvantaged due to poor modeling of the rest of the visual world (Torralba & Efros, 2011).

For example, in the context of email classification, if the training dataset includes a higher proportion of spam emails than non-spam emails, the machine learning model may be more sensitive to spam emails and may classify a higher proportion of non-spam emails as spam (Zhou et al., 2014).

Negative set bias can be mitigated by balancing the training dataset or weighting the training data to give greater importance to positive examples (Chawla, 2005).

2.2.8. Problem Framing bias

Problem framing errors can also cause bias (Srinivasan & Chander, 2021). For instance, if a credit card company aims at predicting customer trustability using AI, the concept of creditworthiness must be well-defined and estimated. However, "creditworthiness" is a rather vague concept (Barocas & Selbst, 2016). Problem framing strongly depends on the company's goals: maximizing the profit margin or maximizing the number of repaid loans.

However, as Solon Barocas, an assistant professor at Cornell University who specializes in fairness in machine learning emphasizes, "those decisions are made for various business reasons other than fairness or discrimination" (Hao, 2019). If the algorithm discovered that granting subprime loans lead to profit maximization, it would eventually lead to predatory behavior, even if it was not the intention of the company.

2.2.9. Recent Bias Mitigation Approaches

It is essential to be aware of these types of data bias and build a versatile mitigation strategy in order to avoid their effects and ensure that machine learning models are accurate and reliable (Baeza-Yates & Ribeiro-Neto, 2011). In order to address data bias in machine learning, it is recommended to use diverse and representative datasets, apply statistical techniques to adjust for bias, and use multiple methods to validate results (Suresh & Guttag, 2019).

Additionally, the use of human-in-the-loop approaches, where a human is involved in the decision-making process, can also help to mitigate bias in AI systems (Xin et al., 2018). However, some studies warn that systems with one or too few human experts are insufficient due to human agent's bias. One solution to it might be a hybrid pipeline with multiple human experts and a classifier to share the decision making load and reduce bias (Keswani et al., 2022).

It is also important to be transparent about the data sources and methods used in order to allow for external scrutiny and reproducibility (Baeza-Yates & Ribeiro-Neto, 2011).

3. Digital Ageism in AI systems

The digital age has brought numerous advancements and innovations that have transformed the way we live, work, and communicate. However, these advancements have also led to the emergence of a new form of discrimination known as digital ageism (Hunsaker & Hargittai, 2018). Nowadays, digital ageism is addressed as a critical issue and a global priority by the World Health Organization (WHO) in their annual Global Report on Ageism (World Health Organization, 2022).

Digital ageism refers to the discrimination or prejudice against individuals based on their age or generation in the digital world. Digital ageism can manifest in various ways, such as the exclusion of older individuals from technology training and education, the assumption that older individuals are not capable of using technology, and the creation of age-based stereotypes in the media and advertising (Charles & Carstensen, 2010; Zickuhr & Smith, 2012).

In this chapter, we will examine the various forms of digital ageism and how they impact older individuals in the digital world. We will also discuss the ways in which digital ageism intersects with other forms of bias, such as racial and ethnic bias, and how these intersections can compound and amplify the negative effects on marginalized groups. Finally, we will explore potential solutions for addressing and combating digital ageism in order to create a more inclusive and equitable digital society for all.

3.1. Forms of Digital Ageism

Digital ageism encompasses a range of forms of discrimination, including exclusion from technology training and education, negative stereotypes and prejudices, and lack of accessibility of technology for older adults.

One common form of digital ageism is the exclusion of older individuals from technology training and education (Czaja et al., 2008). This can occur when older individuals are not offered the same opportunities for technology training and education as their younger counterparts, leading to a lack of digital literacy and skills among older adults. The consequences of exclusion from technology can be significant for older adults: from limited access to job opportunities and social connections, to contribution to social isolation (Hultsch et al., 1999).

Another form of digital ageism is the assumption that older individuals are not capable of using technology (Choi et al., 2020; Palmore, 2001). Such a stereotype can lead to older individuals being excluded from certain technological platforms and experiences, or being treated with condescension when attempting to use technology. Older individuals who rely on technology can be particularly disadvantaged in their daily activities, such as staying in touch with loved ones or managing their health.

Older adults often face discrimination during the design process of digital technologies. Such evidence is presented in a recent study that analyzed 7 facial image datasets. Age discrimination was manifested in the labeling of the datasets, where extensive age intervals were assigned to older adults in datasets (Chu et al., 2022). For instance, groups for participants of younger age were categorized into narrow age groups within each dataset, such as 13 to 19, and 20 to 36 years old, compared to a considerably more pervasive category 60+ or 66+ years old, despite decades of physical and mental changes for those individuals.

Digital ageism can also manifest in the form of age-based stereotypes and prejudices in the media and advertising (de Paula Couto & Wentura, 2017). For example, older individuals may be depicted as out-of-touch or unable to keep up with new technologies, leading to negative stereotypes that can further exclude them from participating in the digital world (de Paula Couto & Wentura, 2017).

Overall, there is doubtfully enough data to represent older individuals. Essentially, the existing data also fails to include and depict healthy ageing, underrepresenting older adults' needs, interests, and aspirations, which confirms ageist stereotypes (Chu et al., 2022).

3.2. Intersections of Digital Ageism and Other Forms of Bias

Digital ageism often intersects with other forms of bias, compounding and amplifying the negative effects of bias on marginalized groups (Drydakis et al., 2018; World Health Organization, 2022).

For instance, older individuals from marginalized racial and ethnic groups may face double discrimination due to both their age and their racial or ethnic identity (Drydakis et al., 2018). The study shows that older applicants received a lower number of job interview invitations compared to younger participants. However, the study also states that a study group with people of color as participants had even worse vacancy access. The outcome implies that people with minority ethnicities face a higher level of ageism compared to the majority race representatives (Drydakis et al., 2018).

In addition, digital ageism can intersect with other forms of bias in the development and design of technology. AI and machine learning algorithms that are trained on biased datasets may produce biased outputs that disproportionately negatively impact certain age and racial or ethnic groups (Mehrabi et al., 2021). This can occur in a plethora of contexts, such as in the development of age-based or racially biased advertising or the use of AI in hiring decisions (Mehrabi et al., 2021).

A recent study shows that the intersection of ageism and sexism is a prominent combination even among designers and developers of technologies for older people (Chen & Petrie, 2022). The authors conducted a qualitative study with in-depth interviews with technology designers and developers and found that both male and female participants held negative attitudes toward older workers. In particular, older women were found to face double discrimination due to the intersection of ageism and sexism. This is consistent with previous studies that have shown that women face intersectional discrimination based on their race and gender, in addition to age (Harnois, 2014; Stypińska, 2021).

These intersections of digital ageism and other forms of bias can have significant negative impacts on marginalized groups, such as limiting access to job opportunities and social connections and contributing to social isolation and decreased social capital. It is important to recognize and address these intersections in order to create a more inclusive and equitable digital society for all individuals, regardless of age or identity (Drydakis et al., 2018).

3.3. Potential Solutions

There are several potential solutions for addressing and combating digital ageism. One widely suggested approach is to increase the availability and accessibility of technology training and education for older individuals (Friemel, 2016; Mitzner et al., 2010; Niehaves & Plattfaut, 2014). This solution involves providing targeted technology training programs for older adults, as well as ensuring that these programs are available in a variety of locations and formats to accommodate different learning styles and needs (Mitzner et al., 2010).

Apart from accessibility, a positive user experience (UX) can play a significant role in bringing safety and comfort to older adults as users of various applications. A study shows that the UX in information and communication technologies, poorly adjusted to older individuals' needs and user behavior patterns, distances them from the digital world, causing digital exclusion and, consequently, feeling of loneliness among the participants (Lagacé et al., 2015). A recent study by Chen and Petrie (2022) reaffirms that technology specialists for older people as users should receive adequate de-biasing training in order to reduce the number of biased experts in the field.

Another solution is to challenge and debunk age-based stereotypes and prejudices in the media and advertising (Zickuhr & Smith, 2012). The specific steps can be promoting more positive and accurate portrayals of older individuals in the media, as well as calling out and addressing instances of ageism in advertising and media content.

Overall, there is a need for more inclusive and equitable design and development of technology, including AI and machine learning algorithms. It is crucial to ensure that these systems are trained on diverse and representative datasets, as well as implementing measures to mitigate and address potential biases in the outputs of these systems (Mehrabi et al., 2021).

3.4. Successful Practices and Initiatives Worldwide

It is worth mentioning the initiatives and programs that have been implemented to combat ageist digital inequalities worldwide. For instance, the European Commission's (EAEA) campaign "New skills agenda for Europe" (2019) aims to promote the development of digital skills, including those of older adults, and ensure that they are not left behind in the digital transformation. The "Silver Surfers" program in the UK, launched by Age UK and TalkTalk, provides training and support for older adults to help them acquire the digital skills they need to participate fully in the digital world (Age UK & TalkTalk, 2014).

Similarly, US-based non-profit organizations, such as American Association of Retired Persons (AARP), are focusing on issues affecting older individuals over age fifty. As of 2018, the AARP group reported to have made significant contributions towards improving the lives of over 38 million members, including providing access to better economic security, consumer protection, and healthcare, promoting affordability and quality in long-term care, and fostering the development of livable communities (AARP, n.d.). Their program (Older Adult Technology Services (OATS), 2022), provides resources and support for older adults to learn about and engage with technology, breaking down barriers to digital inclusion.

Initiatives aimed at promoting digital inclusion for older adults are highly relevant to the ageing population in Japan, which is one of the fastest-growing in the world. According to the Annual Report on the Aging Society (2017), dementia is forecasted to have an effect on one in five people in Japan by 2025. Access to digital skills and resources is vital for older individuals in Japan to participate fully in modern society and maintain their quality of life.

There are several notable initiatives in Japan aimed at promoting digital inclusion for older adults. For instance, the

Ministry of Internal Affairs and Communications launched the "Silver Human Resources Center" program in 1974, to support older job seekers (Weiss et al., 2005). Nowadays, Silver Human Resources Center also provides digital literacy training to older adults in Japan. The program aims to create a network of people who can support older adults in learning about and using digital technology.

In addition, the Japanese government has implemented policies to encourage businesses to develop age-friendly technologies and services, such as the "Universal Design" policy, which promotes the design of products and services that are accessible to all, regardless of age or ability (Ministry of Economy, Trade and Industry, 2020).

These initiatives demonstrate the Japanese government's commitment to promoting digital inclusion and addressing digital ageism in Japan, and serve as an important model for other countries to follow. As a result of these programs, unique cases have emerged, such as that of an 83-year-old female app game developer (Government of Japan, 2018) or a restaurant that is run and maintained by people affected by dementia (Government of Japan, 2019).

Raising awareness and a better understanding of digital ageism and its impacts on older individuals can facilitate a shift toward a more inclusive and equitable digital society for all, as well as increase life quality for older individuals.

4. Ethnic Bias in AI Systems

Ethnic bias in AI refers to the tendency for AI systems to produce biased outcomes that disproportionately harm or discriminate against certain racial or ethnic groups. Bias can occur when the data used to train AI systems reflects and reinforces existing societal biases and inequalities. Resolving ethnic bias in AI is vital due to far-reaching consequences for those who are targeted by it, including discrimination, marginalization, and reduced opportunities (Zafar et al., 2017). Additionally, ethnic bias in AI can magnify existing societal inequalities, leading to further harm and injustice (Bolukbasi et al., 2016).

In this chapter, we will examine the ways in which ethnicity-based bias can occur in AI, the consequences of this bias, and efforts to mitigate or eliminate it. We will review relevant studies on the topic and discuss practices for designing and evaluating AI systems to reduce the risk of ethnicity-based bias.

4.1. Impact of Ethnic Bias in AI Systems

The issue of ethnic bias in artificial intelligence has gained increasing attention in recent years. Incidents have exposed the vulnerability of AI to perpetuating existing societal biases, leading to calls for increased efforts to identify and address such biases in order to promote justice and equality for all individuals.

One example of ethnic discrimination in the digital world is the use of targeted advertising by Facebook (now: Meta Platforms). In 2016, it was revealed that Facebook allowed advertisers to exclude African, Hispanic, and other "ethnic affinities" from seeing advertisements (Ali et al., 2019). Such practice magnified existing inequalities and discrimination, as individuals from certain ethnicities may have been disproportionately excluded from seeing certain advertisements based on their zip code or other factors that are correlated with ethnicity. Despite public exposure and repeated media investigations, the problem has remained over years (Ali et al., 2019; Angwin & Parris, 2016; Angwin et al., 2017).

Another study shows that in order to select a look-alike audience, Facebook tries to infer the attributes that distinguish the audience from the general population, recurrently causing representation bias. Such bias distribution might potentially incline biases in a source audience of several thousand to a lookalike audience of tens of millions (Speicher et al., 2018). As a result of this type of discrimination, individuals from certain ethnicities may be disadvantaged in terms of access to job opportunities, housing, and other resources (Dastin, 2018).

Finally, a lack of ethnic and racial diversity has been observed within academic settings (D. Zhang et al., 2021). According to the AI Index 2021 Annual Report, among the new AI PhDs in the USA in 2019, the largest percentage (45.6%) are white representatives (non-Hispanic), followed by Asian representatives (22.4%). In comparison, only 2.4% were African American (non-Hispanic) and 3.2% were of Hispanic ethnicities. Ethnic underrepresentation in AI research and development can limit the diversity of design and deployment of AI systems, highlighting the need for increased efforts to promote diversity and inclusivity in the field.

4.2. Consequences and Implications of Ethnic Bias

Ethnicity-based bias in AI can manifest in a variety of ways, including the amplification of existing societal biases and discrimination (Zafar et al., 2017). Some specific forms of ethnicity-based bias in AI include the issues depicted in this chapter.

4.2.1. Accuracy Disparities in Face Recognition

Accuracy disparities in AI systems can become a problem, as certain ethnicities may be more accurately represented in these systems, leading to unequal treatment depending on the ethnicities of individuals.

To illustrate, AI systems used for facial recognition or language processing may be more accurate for certain ethnicities, resulting in discriminative outcomes for individuals based on their ethnicities (Buolamwini & Gebru, 2018; Caliskan et al., 2017; Dastin, 2018). According to a recent study by MIT and Microsoft, false arrests or incorrect identification of suspects are possible as a result of poor recognition and identification accuracy by AI systems for individuals from certain ethnicities. The experiment by Buolamwini and Gebru (2018) shows that the maximum difference in face recognition error rate between the lighter skin tone male groups and darker skin tone female groups, best and worst classified groups respectively, is 34.4%. This accuracy disparity is largely caused by datasets exceedingly composed of

lighter-skinned participants (79.6% to 86.2%, depending on the dataset).

Additionally, another research has proved that these accuracy disparities can be particularly pronounced for individuals who are members of multiple marginalized groups, such as women of color (Else-Quest & Hyde, 2016).

4.2.2. Hiring and Lending

Implementation of AI in decision-making processes in hiring and lending has sparked concerns about the risk of ethnic bias and the exacerbation of current inequalities.

For instance, an AI system involved in the hiring process may be more likely to shortlist job applicants from certain ethnicities, causing an unfair advantage for individuals from those groups (Caliskan et al., 2017). A recent study showed that word embedding, a popular framework to transform text data into structured vectors that can be more easily processed by a computer, can lead to sexism and other forms of discrimination (Bolukbasi et al., 2016). Word embedding has been used in various machine learning tasks, including AI systems, trained to assist in hiring decisions.

Similarly, an AI system used in lending decisions may be more likely to approve loans for individuals from certain ethnic groups, leading to an unequal distribution of financial opportunities (Barocas & Selbst, 2016; Zafar et al., 2017).

4.2.3. Law Enforcement and Security Contexts

The use of AI systems in law enforcement and security contexts has garnered attention due to the potential for racial profiling and discrimination. AI systems used in these contexts may be more likely to identify individuals from certain ethnicities as potential suspects, leading to false accusations and other acts of discrimination (Caliskan et al., 2017; Else-Quest & Hyde, 2016; O'Neil, 2016).

In particular, there are several ways in which a predictive policing algorithm may impose discrimination on individuals of certain ethnicities. To illustrate, a predictive policing algorithm in Florida, United States, in 2013 and 2014 was more likely to falsely assign high risk scores to individuals of color as potential suspects, with only 20% of the correct prediction rate (Angwin et al., 2022). It is possible that the prediction algorithm was trained on biased data that includes a disproportionate number of individuals from certain ethnicities who have been arrested or convicted (Eubanks, 2018; O'Neil, 2016). If the algorithm is trained on this biased data, it may be more likely to identify individuals from those ethnicities as potential suspects, even if they are no more likely to commit crimes than individuals from other ethnicities.

Another possibility is that the algorithm is using factors correlated with ethnicity, such as zip code or socioeconomic status, as input (Eubanks, 2018). If these factors are correlated with ethnicity and are being used by the algorithm to predict the likelihood of criminal activity, it may be more likely to identify individuals from certain ethnicities as potential suspects, even if their ethnicity is not directly related to their likelihood of committing a crime (O'Neil, 2016).

Biased crime risk prediction systems can jeopardize the human rights of marginalized groups, leading to increased discrimination and social inequality. Therefore, it is crucial to implement strict regulations and conduct thorough checks before implementing these technologies.

4.3. Potential Solutions

Mitigation and elimination of these forms of bias should become vital in the development and deployment of AI systems. This can involve using diverse and representative datasets, implementing fairness and accountability measures, and regularly evaluating AI systems for bias (Dwork et al., 2012).

To address ethnic bias in AI systems, it is important to take a multi-faceted approach that involves both technical and societal solutions. One solution is the use of diverse and representative training data to develop AI algorithms in order to ensure that the algorithms are more representative of the populations they serve (Dwork et al., 2012).

Another solution is conducting regular bias audits of AI systems, and active monitoring of their performance for evidence of bias (Landers & Behrend, 2022). However, such audit and cross-checking imply additional hours of work for current employees, or, potentially, additional hiring. While large companies, such as IBM, can afford it and actively implement these complex techniques(Hobson & Dortch, 2022), startups and companies with exiguous budgets might not possess this opportunity. Alternatively, there are currently tools that help detect and measure bias in models, as well as calculate the bias drift over time (Simon, 2022). One such tool is Amazon SageMaker by Amazon Web Services (AWS).

In addition to these technical solutions, organizations must also prioritize diversity, equity, and inclusion in their hiring practices and organizational culture. This includes creating a diverse team of researchers and engineers and establishing ethical review processes for AI systems development (Floridi, 2019). A similar to suggested set of practices and tools is implemented within Re:work Unbiasing course, an open-source educational course by Google (2017). The course aims at reducing potential unconscious bias for hiring and promotion decisions by providing a set of practices, checklists, facilitator guides, and team discussion guides that can be adjusted to the user's team.

Implementing these solutions requires a commitment to transparency, accountability, and continuous improvement. This includes regularly reporting on the performance of AI systems and publishing data on their biases and limitations. Additionally, it is crucial for organizations to prioritize the development of ethics and governance frameworks to ensure the responsible development and deployment of AI systems (Floridi, 2019).

Addressing ethnic bias in AI systems is a complex and ongoing challenge. However, by taking a diversified approach, communities can help ensure that AI is developed and used in a way that is equitable, just, and beneficial for all.

5. Methodology

In this study, a controlled experiment was conducted to investigate the potential for biased results for different ethnic groups in ASR systems. The main objective of the study was to evaluate the performance of commonly used ASR systems in recognizing speech from non-native proficient English speakers of various ethnic groups, and put these results in comparison with canonical examples of American, Australian, and British native English speakers.

The methodology for this experiment included the selection of a sample of English-speaking participants representing a range of ethnicities, and the recording of speech samples from each participant. The speech samples were then manually checked in terms of quality standards fulfillment, and finally processed through the selected ASR systems. The ASR Systems selected for the study are as follows: Gboard for Android by Google, Apple keyboard for iOS, and Whisper by Open AI. The recognition accuracy was measured and compared across ethnic groups.

The study will provided insights into the potential biases in AI algorithms and their impact on different ethnic groups, which is discussed in the Discussion chapter of the thesis.

The experiment consists of four phases: Planning, Performing, Reviewing, and Closing stages, as shown in Figure 1.

The results of the study were analyzed using statistical methods to determine whether there are any significant differences in the recognition accuracy of the ASR systems for different ethnic groups.

This research aimed to contribute to the understanding of potential biases in AI algorithms and their impact on different ethnic groups and to identify possible solutions to mitigate such biases.

5.1. Automatic Speech Recognition Applications

According to recent data, the utilization of AI transcription applications has grown exponentially in recent years. Research has shown that the integration of AI in transcription applications has the potential to improve communication and accessibility for individuals with speech or language impairments (J. Zhang et al., 2023). The rapid growth in the use of these applications highlights the increasing importance of AI in this field.

Among the most widely used applications in this category are Gboard by Google, and the Apple keyboard, both are pre-installed on a large number of mobile devices nowadays. In contrast, another ASR selected for the experiment is a relatively new product picked to represent the state of the art of ASR solutions. Whisper by OpenAI has been described as "revolutionary" in the field of AI transcription by experts in the field (Ansari, 2022).

Gboard has emerged as a leading player in the field of AI transcription, with over 5 billion downloads reported in 2022 (Google Play, n.d.). The app is generally installed out-of-the-box on many Android-based mobile devices and supports over 900 languages, as per the application's description. Gboard is also available on the iOS operating system and can be installed on iOS-based devices through the App Store. However, both the number of downloads and supported languages are significantly lower on the iOS App Store (App Store, 2016).

The Apple iOS keyboard is a default application that is pre-installed on the iOS operating system. According to Tim Cook, Chief Executive Officer of Apple, there were at least 1.65 billion Apple devices, as of January 2021 (Nellis, 2021). This suggests a widespread availability and usage of the Apple iOS keyboard among individuals who utilize Apple devices.

The third voice transcription application under examination is Whisper, an open-source state-of-the-art ASR system. The system has been trained on a vast amount of supervised data, precisely, 680,000 hours of data according to Radford et al. (2022). Whisper is a research project developed by OpenAI, a leading AI research and deployment research company, based in California, USA. OpenAI was founded by notable figures in the technology industry, including Elon Musk and Sam Altman (OpenAI, 2022).

In contrast to Gboard and iOS virtual keyboards, is a large ASR model executed in clusters of servers in the cloud. As such, It is not meant for on-device real-time transcription like the other products surveyed. The ASR system takes an average of 25 to 140 seconds to transcribe a voice sample. This feature may be useful for more complex transcription tasks that require higher accuracy and can tolerate either long processing time or increased execution costs.

It is interesting to note that, while Gboard does not include punctuation in its transcription, Apple adds periods and commas based on pauses between words, and Whisper accurately places periods at the end of each sentence.

5.2. Participants and Data Collection

There are canonical samples of native speakers' recordings for Harvard sentences located in the Open Speech Repository and further used for comparison with non-native English-speaking participants' transcription data (Open Speech Repository, n.d.).

The study population included 210 participants, of which 23 participants' recordings were excluded from the experiment due to low audio quality or reading mistakes. The 187 recordings were collected from proficient and fluent English speakers of 55 different ethnicities. The 187 participants (99 male and 88 female) were divided into 4 ethnic groups: African, Asian, European, and Hispanic/Latin, with each group containing 29, 60, 68, and 30 samples respectively.

As a prerequisite for participation, it was required that all individuals possess proficient English language skills, which were demonstrated through either their enrollment in English-based academic programs or their utilization of English within their professional settings.

The average age of participants in the study was calculated to be 28.62 years old. The study had originally aimed to

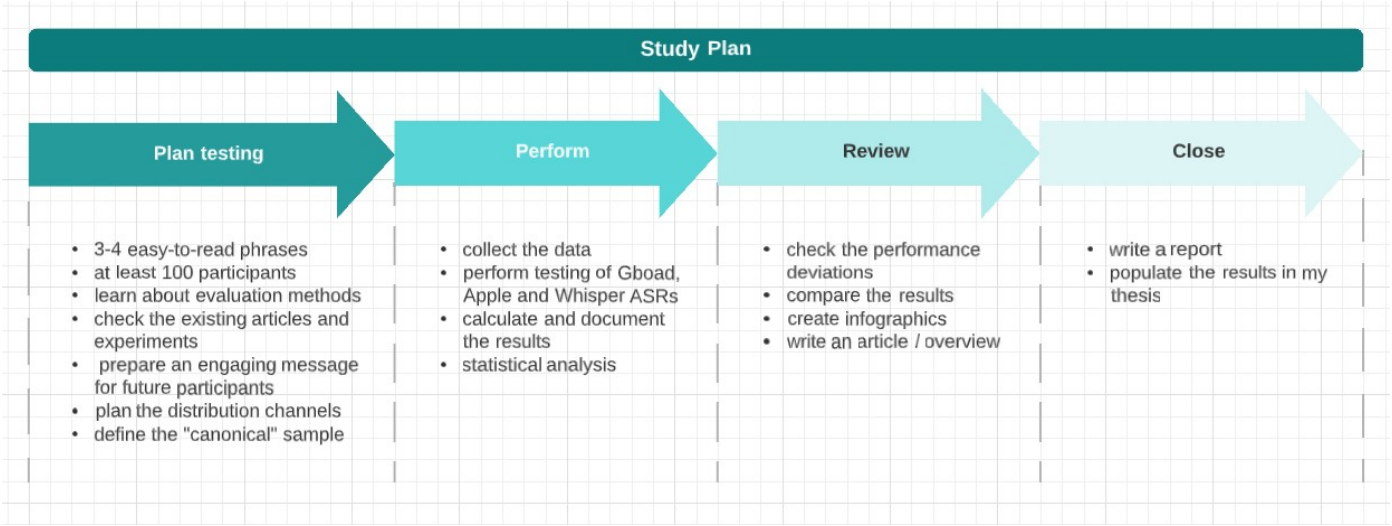

Figure 1: Experiment Planning and Implementation Phases

investigate age-based bias and therefore included the tracking of participants' ages. However, it was found that the majority of participants fell within the age range of 20-30 years. As a result of the limited representation of older individuals, the decision was made to focus on ethnic bias analysis instead.

The following channels were used for the data collection:

- Technical University of Munich student community;
- Tokyo Institute of Technology student community;
- AIESEC organization channels;
- Social Media (LinkedIn, Instagram);
- TUM SEED Center internal channels;
- Friends and acquaintance contacts.

The study was conducted in accordance with ethical considerations. Informed consent was obtained from all participants, and the data was collected and handled in a way to ensure the anonymity and confidentiality of the participants.

5.3. Harvard Sentences

In this study, it was hypothesized that the performance of ASR systems might be influenced by the ethnicity (accent) of a speaker. This was based on the observation that many ASR systems are trained on speech samples from a limited variety of dialects or pronunciation variances (Dahl et al., 2012; Li et al., 2018). In order to investigate this hypothesis, an experimental design was implemented to evaluate the extent to which different accents affected the recognition level of the most commonly used ASR systems. Participants of different ethnicities and accents were asked to speak the same set of phrases from the Harvard sentences.

Harvard sentences are a set of standardized phrases (720 sentences). Developed in the 1950s by Harvard researchers, they are still widely used to test everything from cell phones to Voice over Internet Protocol (VoIP). According to David Pisone, director of the Speech Research Laboratory at Indiana University, Harvard Sentences have become the golden standard for speech-to-text engineers and speech scientists (S. Zhang, 2015).

There are other standardized sets of words and phrases for testing and training ASR, but Harvard sentences are among the oldest and most popular. The sets of sentences are involved in numerous experiments and research in the past, (Schwab et al., 1985) and recent years (Loebach et al., 2010; Smith et al., 2019). The set of phrases within Harvard sentences selected for the study is set H5, sentences 1, 2, and 3 (Harvard Sentences, n.d.).

1. *"A king ruled the state in the early days."*
2. *"The ship was torn apart on the sharp reef."*
3. *"Sickness kept him home the third week*."

The speech samples were analyzed for recognition accuracy and errors. This design allowed for a direct comparison of the ASR systems' performances across different accents, providing insight into the potential influence of accents on recognition level.

5.4. Data Analysis

This sub-chapter presents the methods and results of the data analysis conducted on the results of the experiment, including the implementation of the Word Error Rate (WER) algorithm, Power analysis, Analysis of Variance (ANOVA), and post-hoc Tukey Test.

5.4.1. Word Error Rate Computation

Speech recognition research typically evaluates and compares systems based on the word error rate (WER) metric (Radford et al., 2022). In particular, the performance of ASR systems is measured by computing the WER algorithm for the

transcribed data. Word error rate is a metric that is based on string edit distance. It identifies all differences between the model's output and the canonical sample transcription (Park et al., 2008). The formula for WER is as follows:

$$WER = \frac{S + D + I}{N}$$

where S, D, and I are the number of substitutions, deletions, and insertions respectively, and N is the total number of words in the reference transcription.

In order to minimize human errors and avoid miscalculations, data evaluation was automated. The code was written in JavaScript and used the Google Sheets API to compute the WER for a set of voice transcription data. The key functions of the code are presented and explained below, while the full version of the code is available in *Appendix D. Word Error Rate Implementation* file.

```
function computeErrorRates() {
  const sheet = SpreadsheetApp.
      getActiveSpreadsheet().getSheetByName('Data
      ');
  const canonical = sheet.getRange('K3').
      getValues()[0][0];
  const gboard = sheet.getRange('E3:E192').
      getValues();
  sheet.getRange('D3:D192').setValues(process(
      gboard, canonical));
  const apple = sheet.getRange('G3:G192').
      getValues();
  sheet.getRange('F3:F192').setValues(process(
      apple, canonical));
  const whisper = sheet.getRange('I3:I192').
      getValues();
  sheet.getRange('H3:H192').setValues(process(
      whisper, canonical));
}

function process(entries, expected) {
  return entries.map((entry) => {
    const answer = wer(entry[0], expected);
    return [(answer * 1000 | 0) / 1000];
  });
}

function trim(text, chars) {
  let low = 0;
  for (; low < text.length && chars.includes(text
      [low]); low++);
  let high = text.length - 1;
  for (; high >= low && chars.includes(text[high
      ]); high--);
  return text.slice(low, high + 1);
}

function cleanup(text) {
  return text
  .split(" ")
  .map((s) => trim(s.trim(), ".,_").toLowerCase()
      )
  .filter((s) => s !== "");
}

function wer(text, expected) {
  text = cleanup(text);
  expected = cleanup(expected);
  const n = text.length;
  const m = expected.length;
  const dp = Array.from({ length: n + 1 }, (_, i)
      =>
  Array.from({ length: m + 1 }, (_, j) => (i ===
      0 ? j : j === 0 ? i : 0)) );
  for (let i = 1; i <= n; i++) {
    for (let j = 1; j <= m; j++) {
      dp[i][j] =
      text[i - 1] === expected[j - 1]
      ? dp[i - 1][j - 1]
      : 1 +
      Math.min(Math.min(dp[i - 1][j], dp[i][j -
          1]), dp[i - 1][j - 1]);
    }
  }
  return dp[n][m] / m;
}
```

The code is composed of several functions that work together to calculate the WER.

1. computeErrorRates(): This function is the main function that is called to initiate the computation of word error rates (WER) for the data collected in the experiment. The function starts by getting the active spreadsheet and searching for a sheet named "Data".
2. process(entries, expected): This function takes two inputs, "entries" and "expected", and maps the entries to their corresponding WER values by calling the wer(entry[0], expected) function and returning an array of the WER values truncated to three decimal digits of precision.
3. trim(text, chars): This function is used to remove unwanted characters in the beginning or the end of a string of text. It takes two inputs, "text" and "chars", where "text" is the string of text and "chars" is a string of characters to remove.
4. cleanup(text): This function is used to clean and prepare the text for the WER calculation. It takes one input, "text", and performs several operations such as splitting the text into words, removing unwanted characters and converting all characters to lowercase.
5. wer(text, expected): This function calculates the WER between the given text and the expected text. It takes two inputs, "text" and "expected", and implements a dynamic programming algorithm to find the minimum number of edits required to transform the text into the expected text. The function returns the WER value.

The code first calls the computeErrorRates() function, which initiates the process of computing the WER values for the data. It then uses the getRange() method to get the data from the sheet, and passes it to the process() function along with the canonical text. The process() function then maps the data to their corresponding WER values by calling the wer() function, and sets the values back.

The results for WER computations are presented in the *Appendix A. Voice Transcription Experiment Data* file under the sheet "*Data*".

5.4.2. Power Analysis and ANOVA

In order to determine the statistical significance of the differences in ASR systems' performance between different ethnic groups, we performed Power analysis, ANOVA One-Way with Unequal n's, and post-hoc Honest Significant Difference (HSD) Tukey Test on the results of the experiment.

The power analysis, conducted using the NQuery software, was used to evaluate the effect size and ensure that the experiment had sufficient power to detect meaningful differences between groups (O'Brien & Muller, 1993). The one-way ANOVA, which is a statistical test that compares the means of multiple groups, was conducted using the Excel extension XLMiner Analysis tool pack. This test was used to determine if there were significant differences in ASR performance between the four ethnic groups (African, Asian, European, and Hispanic/Latin) in our study.

In case of significant p-value ANOVA results, we also performed a post-hoc Tukey test in order to compare the pairs of ethnic group performances to each other, and see where the differences lie between the groups (Copenhaver & Holland, 1988; Gleason, 1999; "R: The Studentized Range Distribution", n.d.). Detailed report with the results, tools, and additional files is included in *Appendix B. Anova & NQuery Power & post-hoc Tukey Results* file.

In this study, we accept that the Null Hypothesis is the following: "ASR systems' performance is equally good for all ethnicity groups".

6. Results

In this chapter, we present the results of the experiment examining the performance of ASR systems Gboard, Apple keyboard and Whisper for different ethnic groups. The performance of the ASR systems was measured using the word error rate (WER) algorithm. The results were analyzed using power analysis, one-way ANOVA, and post-hoc Tukey Test.

The results are presented in terms of error rates, tables, and figures, which show the performance of the ASR systems for each ethnic group, as well as the comparison between the ethnic groups.

6.1. Data Analysis Results

The results of the ANOVA One-Way with Unequal n's analysis indicate that there is a statistically significant difference in the performance of the three ASR systems, Gboard, Apple, and Whisper, with respect to recognizing speech from different ethnic groups. The analysis was conducted on 4 groups of participants representing African, Asian, European, and Hispanic/Latin ethnicities. The analysis results can be found in Table 1.

The test significance level, α, which is the probability of rejecting the null hypothesis when it is true, was set at 0.05 for each system. The number of ethnic groups for each tool is 4, the total sample size is 187 participants.

Additionally, the N as multiple of n_1, $\sum r_i$, is 6.448275862 for all systems which indicates that the sample size of each ethnic group is relatively similar for each system which is favorable for comparing the results across the groups and ASR systems.

The variance of means, V, was found to be 0.0034537944 for Gboard, 0.0017757824 for Apple, and 0.0038377534 for Whisper. These results propose that there is a relatively small difference between the means of the scores for the different ethnic groups for ASR Apple, compared to ASR GBoard and ASR Whisper.

The common standard deviation, σ, which is the square root of the variance, was found to be 0.1847353293 for Gboard, 0.2742167753 for Apple, and 0.2175543915 for Whisper. This indicates that the data points are further spread out for Apple than for Gboard and Whisper.

A higher effect size, Δ^2, indicates a larger difference between the groups. This implies that the effect of the ethnic groups on the performance of Apple is lower than the effect of the ethnic groups on the performance of Gboard and Whisper.

Power, the probability of detecting a true effect if one exists, was found to be 96.4% for Gboard, 38.5% for Apple, and 91.4% for Whisper. This suggests that the Gboard and Whisper surveys have a high probability of detecting a true difference if one exists, whereas Apple's power is lower. It is worth noting that the power for Apple is lower than desired ($P_3 \leq 80\%$), which is an indication that the sample size for this system should be increased in future studies.

In order to test the null hypothesis, we perform the One-Way Analysis of Variance (ANOVA) on the tool NQuery for each of the three tools.

The results of the one-way ANOVA for Gboard (Table 2) state that there is a statistically significant difference in the performance of automatic speech recognition systems across different ethnic groups, $F(3, 183) = 3.106$, $p = 0.028$. The null hypothesis, which states that the performance of ASR systems is equally good for all ethnicity groups, is rejected due to the fact that p-value is less than the significance level of 0.05.

The p-value also indicates that there is a 2.8% chance that the observed difference in the performance of Gboard across the ethnic groups is due to random chance, which is considered low and suggests that the difference is likely not due to random variation, but rather a real effect.

As shown in Table 3, the high p-value of 0.481 indicates that there is a 48.1% chance that the observed difference in the performance of Apple keyboard across the ethnic groups is due to random chance, rather than a real effect. This implies that the null hypothesis that the tool (Apple keyboard) has equal performance across ethnic groups cannot be rejected.

The Tukey post-hoc test results for Apple keyboard show that there is no significant difference between the performance of the ASR system for any of the four ethnic groups tested (African, Asian, European, and Hispanic/Latin) (see Table 4). This is indicated by the high p-values for all pairwise comparisons and the inference of "insignificant" for all of them.

Table 1: NQuery One-Way ANOVA with Unequal n's Analysis

	Gboard	***Apple***	***Whisper***
Test Significance Level, α	0.05	0.05	0.05
Number of Groups, G	4	4	4
Variance of Means, V	0.0034537944	0.0017757824	0.0038377534
Common Standard Deviation, σ	0.1847353293	0.2742167753	0.2175543915
Effect Size, $\Delta^2 = V / \sigma^2$	0.101203743	0.0236157482	0.0810851580
Power (%)	**96.39940714**	**38.53636251**	**91.39819872**
N as Multiple of n_1, $\sum r_i = \sum n_i / n_1$	6.448275862	6.448275862	6.448275862
Total Sample Size, N	187	187	187

Table 2: NQuery One-Way ANOVA Results for ASR Gboard

Source of Variation	*SS*	*df*	*MS*	*F*	***P-value***	*F crit*
Between Groups	*0.3179468*	*3*	*0.10598229*	*3.10551347*	***0.02784065***	*2.65396473*
Within Groups	6.2452669	183	0.03412714			
Total	6.5632138	186				

Table 3: NQuery One-Way ANOVA Results for ASR Apple

Source of Variation	*SS*	*df*	*MS*	*F*	***P-value***	*F crit*
Between Groups	*0.1861825*	*3*	*0.06206083*	*0.82530756*	***0.48146338***	*2.65396473*
Within Groups	13.761091	183	0.07519721			
Total	13.947273	186				

Table 4: Post-hoc Tukey Results for ASR Apple

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	0.0099	0.8999947	insignificant
A vs C	0.8398	0.8999947	insignificant
A vs D	1.7755	0.5815922	insignificant
B vs C	1.0389	0.8730357	insignificant
B vs D	2.0577	0.4679167	insignificant
C vs D	1.2597	0.7856598	insignificant

Note: A - African; B - Asian; C - European; D - Latin / Hispanic

The Q statistic, which measures the difference between the means of each pair of groups, is also relatively low for all pairs, further supporting the conclusion of no significant difference in performance across ethnic groups for this ASR system.

The p-value for ASR system Whisper of 0.002 indicates that there is a 0.2% chance that the observed difference in the performance of the ASR system of Whisper across the ethnic groups is due to random chance, which is considered very low and suggests that the difference is likely due to a real effect (Table 5).

The F-value of 5.07 also indicates a significant difference in performance across ethnic groups. This suggests that the null hypothesis that the ASR system Whisper has equal performance across ethnic groups can be rejected.

Overall, the results of the one-way ANOVA test for Whis-

Table 5: NQuery One-Way ANOVA Results for ASR Whisper

Source of Variation	*SS*	*df*	*MS*	*F*	***P-value***	*F crit*
Between Groups	*0.7176599*	*3*	*0.23921996*	*5.06712354*	***0.00214913***	*2.65396473*
Within Groups	8.6394684	183	0.04721021			
Total	9.3571283	186				

per suggest that the null hypothesis can be refuted and the performance of the ASR system likely differs according to ethnic groups.

6.2. Transcription Results

Based on the results presented in Figure 2 and Figure 3, there are significant differences between the performance of ASR algorithms in transcribing the voices of non-native and native English speakers. Non-native speakers had lower transcription rates and higher error rates than native speakers across all ASR algorithms. The highest correct transcription rate was achieved using Whisper by OpenAI (0.866), followed by Gboard by Google (0.778), while the lowest was achieved using Apple keyboard (0.578). On average, the error rate was the highest for Apple keyboard (0.422), followed by Gboard by Google (0.222), while the lowest error rate was achieved using Whisper by OpenAI (0.134).

In contrast, Figure 3 shows that native English speakers had almost perfect transcription rates across all ASR algorithms. The highest transcription rates were achieved using Gboard by Google and Apple keyboard, both scoring a perfect 1.00, while the lowest was achieved using Whisper by OpenAI (0.979) with an insignificant error rate of 0.022.

Overall, the results suggest that ASR algorithms are significantly more accurate at transcribing the speech of native English speakers than non-native English speakers.

When considering the performance of each ASR tool for different ethnic groups, there are also notable differences. Across all ASR tools, Latin/Hispanic speakers had the highest correct transcription rate, while African speakers had the lowest. For Whisper and Gboard, there was a clear trend of increasing performance with Latin/Hispanic and European ethnicities, while for Apple, the performance was relatively consistent across all ethnic groups.

In terms of overall performance, Whisper by OpenAI (see Figure 4) appears to be the most accurate ASR tool with an average correct transcription rate of 86.6% and an average error rate of 13.4%.

Gboard by Google (see Figure 5) also performed relatively well with an average correct transcription rate of 77.8% and an average error rate of 22.2%. However, as shown on Figure 6, Apple keyboard had the lowest performance with an average correct transcription rate of 57.8% and an average error rate of 42.2%.

It is also worth noting that the error rates for Asian and African speakers were consistently lower than those for European and Latin/Hispanic speakers across all ASR tools. Additionally, Whisper by OpenAI appears to have the best performance overall, while the Apple keyboard has the worst performance for non-native speakers.

Overall, the data suggests there are possible biases in the performance of ASR tools, with some ethnic groups experiencing significantly higher error rates than others. It also highlights the need for continued research and development in this area to ensure that AI algorithms are designed and trained in a way that is fair and unbiased for all users.

7. Discussion

The purpose of this thesis was to investigate the hypothesis that the performance of ASR systems may be influenced by the ethnicity (accent) of participants.

The results of the ANOVA analysis show that the null hypothesis is rejected for Gboard and Whisper ASR systems, suggesting that ethnicity has an impact on the performance of these systems. The study showed that there is a statistically significant difference in the performance of ASR systems Gboard, Apple keyboard, and Whisper for participants representing different ethnic groups (African, Asian, European, and Hispanic/Latin).

The results of the ANOVA for the Apple ASR system indicated that the null hypothesis cannot be rejected. However, it is important to note that this system performed exceptionally well for the native English speakers group, implying that there may be issues for non-native English speakers who are not well-represented in the dataset. This highlights the importance of collecting data from a diverse range of participants among non-native English speakers, to ensure that AI algorithms are developed and optimized for a range of users.

The research questions addressed in this thesis were centered on the impact of ethnicity on the performance of ASR systems. The literature analysis suggests that even training and awareness of de-biasing techniques may not be sufficient in eliminating biases in AI algorithms. There are various ways that bias can manifest, and multiple parties are involved in the development and implementation of AI systems.

To further address the issue of bias in AI algorithms, it may be useful to introduce bias review as a mandatory step in the development process. This could involve a team of experts in bias analysis who could provide guidelines or recommendations to avoid possible biases in the design and development of AI algorithms. Such a review process could reduce the burden on individual engineers or developers and make the process more secure. However, further research is

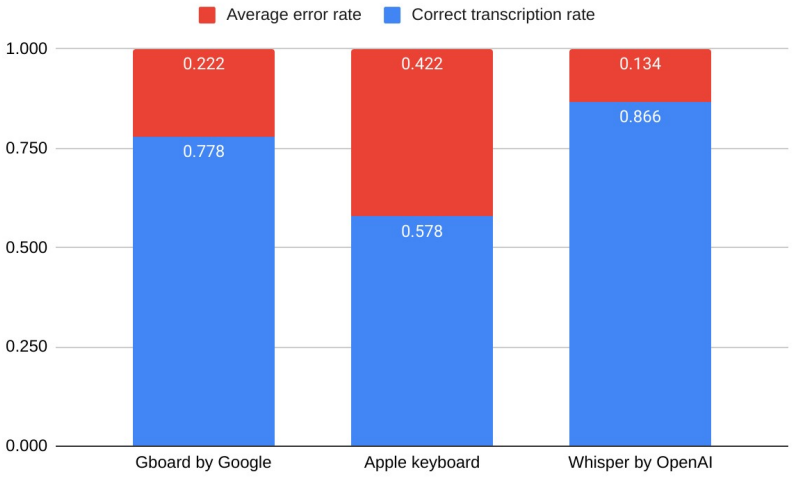

Figure 2: Transcription results for non-native English speakers

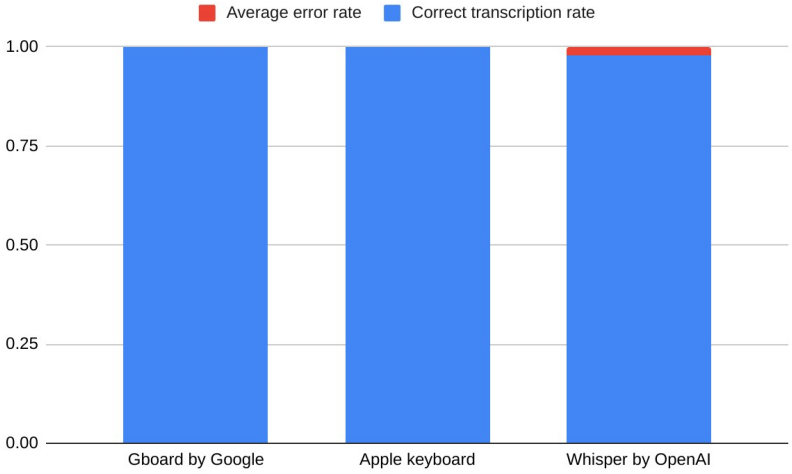

Figure 3: Transcription results for native English speakers

Note: We used canonical Harvard Sentences transcription samples from Open Speech Repository

needed to evaluate the effectiveness of bias review as a tool for reducing biases in AI algorithms.

Additionally, it is crucial to implement strict regulations and conduct thorough checks and inspections before implementing AI technologies.

The findings of this study have important practical implications for the development and use of ASR systems. However, there are certain limitations to this work. The present experiment only focuses on the transcription of audio recordings. It would be more informative to also test the systems on live speech to see if the results generalize to this setting.

As it was discussed previously, not all aspects of the experiment meet the minimum threshold for power, specifically, Apple keyboard, which results in certain findings to be interpreted with caution.

Potentially, further studies will be required with a focus on a more diverse set of phrases in order to evaluate the applications' performance across a wider range of speech patterns and accents, which would provide a more accurate representation of the ASR applications' capabilities. Additionally, it would be more informative for underrepresented groups in the training data sets, where the system may not have seen enough examples to generalize well.

Future research could also explore the specific factors that contribute to the performance differences observed between ASR systems and ethnic groups. This could include exploring the role of language proficiency and exposure. Additionally, further research could investigate ways of improving the accuracy of ASR systems for non-native speakers, such as through accent adaptation or the use of more diverse training data.

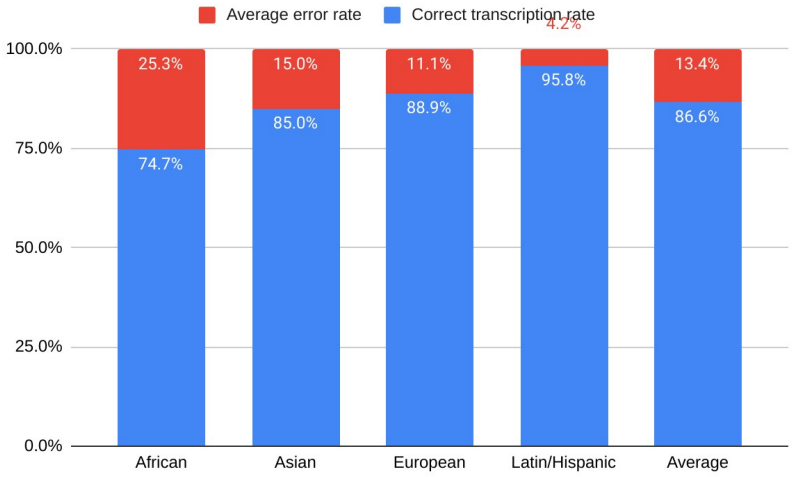

Figure 4: ASR Whisper transcription results

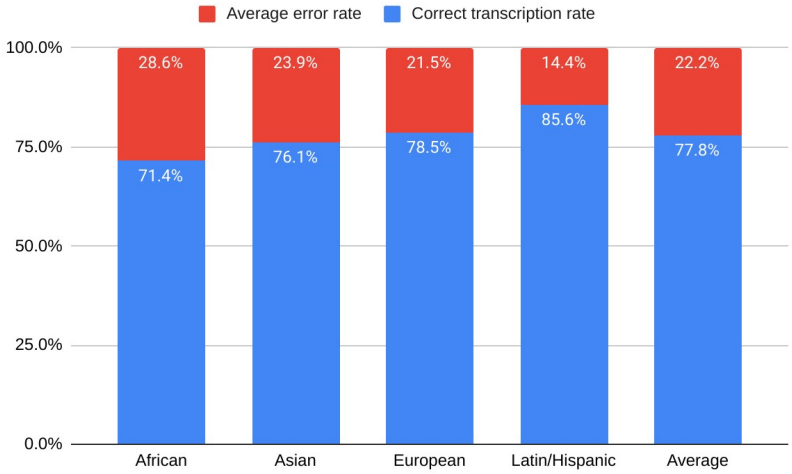

Figure 5: ASR Gboard transcription results

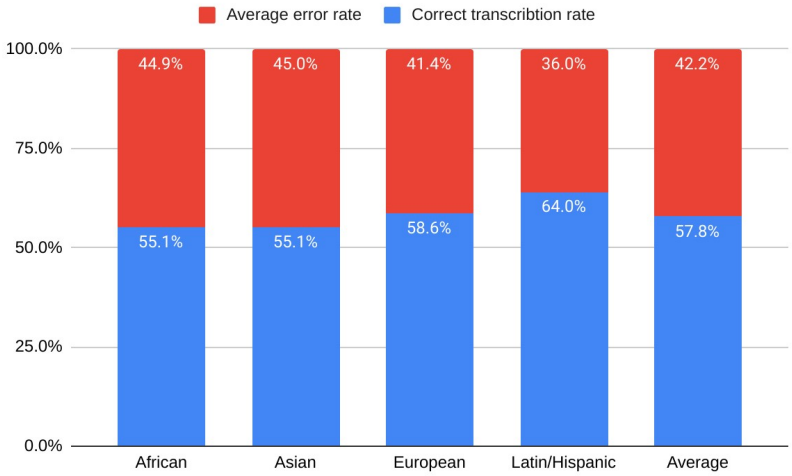

Figure 6: ASR Apple transcription results

Overall, the findings of this thesis suggest that ethnicity (accent) has a significant effect on the performance of ASR systems. It is therefore important for the engineers of AI and ASR systems to be aware of this issue and to take steps to ensure that these systems are not biased toward certain ethnic groups.

8. Conclusion

In conclusion, the results of this thesis demonstrate that ethnicity has a significant impact on the performance of ASR systems. The study provides compelling evidence that ASR systems may not be equally accurate for all users, depending on their ethnicity, which has important practical implications for ASR system development and use. Moving forward, it is important for companies to ensure that their systems are not biased towards certain ethnic groups, and for future research to investigate ways of improving the accuracy of ASR systems for speakers with different accents from all ethnic backgrounds.

While the results of this thesis provide insights into the potential biases of AI algorithms, it is important to note that more data is needed to draw more robust conclusions. Future studies should consider collecting data from a larger and more diverse sample of participants to further investigate the impact of ethnicity and accent on the performance of ASR systems.

References

AARP. (n.d.). AARP Public Policies. Retrieved February 8, 2023, from https://www.aarp.org/about-aarp/policies/?intcmp=FTR-LINKS-INFO-WAS-EWHERE

Age UK & TalkTalk. (2014). Age UK: Helping you love later life - Silversurfers. Retrieved February 7, 2023, from https://www.silversurfers.com/best-of-the-web/lifestyle-best-of-the-web/age-uk-helping-love-later-life/

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on human-computer interaction*, *3*, 199–. http://arxiv.org/pdf/1904.02095

Angwin, J., & Parris, T. (2016). Facebook lets advertisers exclude users by race. Retrieved December 29, 2022, from https://www.cnbc.com/2016/10/28/facebook-lets-advertisers-exclude-users-by-race.html

Angwin, J., Tobin, A., & Varner, M. (2017). Facebook (Still) Letting Housing Advertisers Exclude Users by Race. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Angwin, J., Larson, J. A., Mattu, S., & Kirchner, L. (2022). Machine Bias. *ProPublica*, 254–264. https://doi.org/10.1201/9781003278290-37

Ansari, T. (2022). OpenAI's Whisper is Revolutionary but (Little) Flawed. Retrieved January 17, 2023, from https://analyticsindiamag.com/openais-whisper-is-revolutionary-but-little-flawed/

App Store. (2016). Gboard – the Google Keyboard. Retrieved January 17, 2022, from https://apps.apple.com/us/app/gboard-the-google-keyboard/id1091700242

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54–61. https://doi.org/10.1145/3209581

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology Behind Search* (2nd ed.). Pearson Education. https://archive.org/details/modern-information-retrieval_202102/page/21/mode/2up

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–. https://doi.org/10.15779/z38bg31

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Neural Information Processing Systems*, *29*, 4356–4364. https://arxiv.org/pdf/1607.06520

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, and Strategies*. Oxford University Press. https://books.google.de/books?id=7_H8AwAAQBAJ&printsec=frontcover#v=onepage&q&f=false

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Cabinet Office Japan. (2017). Annual Report on the Aging Society: 2017 (Summary). https://www8.cao.go.jp/kourei/english/annualreport/2017/2017pdf_e.html

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Carvalho, D. T., Pereira, E., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Charles, S. T., & Carstensen, L. L. (2010). Social and Emotional Aging. *Annual Review of Psychology*, *61*(1), 383–409. https://doi.org/10.1146/annurev.psych.093008.100448

Chawla, N. V. (2005). Data Mining for Imbalanced Datasets: An Overview. *Springer eBooks*, 875–886. https://doi.org/10.1007/978-0-387-09823-4_45

Chen, Y., & Petrie, H. (2022). Ageism and Sexism Amongst Young Technicians and Older People in China. *Lecture Notes in Computer Science*, 511–516. https://doi.org/10.1007/978-3-031-08645-8_60

Choi, E. H., Kim, Y. D., Chipalo, E., & Lee, H. C. (2020). Does Perceived Ageism Widen the Digital Divide? And Does It Vary by Gender? *Gerontologist*, *60*(7), 1213–1223. https://doi.org/10.1093/geront/gnaa066

Chu, C., Nyrup, R., Donato-Woodger, S., Leslie, K., Khan, S., Bernett, C., & Grenier, A. (2022). Examining the technology-mediated cycles of injustice that contribute to digital ageism: Advancing the conceptualization of digital ageism: evidence and implications. *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 545–551. https://doi.org/10.1145/3529190.3534765

Copenhaver, M. D., & Holland, B. (1988). Computation of the distribution of the maximum studentized range statistic with application to multiple significance testing of simple effects. *Journal of Statistical Computation and Simulation*, *30*(1), 1–15. https://doi.org/10.1080/00949658808811082

Czaja, S. J., Lee, C. C., Nair, S. N., & Sharit, J. (2008). Older Adults and Technology Adoption. *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, *52*(2), 139–143. https://doi.org/10.1177/154193120805200201

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 30–42. https://doi.org/10.1109/tasl.2011.2134090

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology*, *40*(5), 642–658. https://doi.org/10.1016/j.jesp.2004.02.003

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved December 20, 2022, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

de Paula Couto, M., & Wentura, D. (2017). *Implicit Ageism: Ageism, stereotyping and prejudice against older persons*. The MIT Press. https://doi.org/10.7551/mitpress/10679.003.0006

Dovidio, J. F. (2001). On the Nature of Contemporary Prejudice: The Third Wave. *Journal of Social Issues*, *57*(4), 829–849. https://doi.org/10.1111/0022-4537.00244

Drydakis, N., Macdonald, P. S., Chiotis, V., & Somers, L. (2018). Age discrimination in the UK labour market. Does race moderate ageism? An experimental investigation. *Applied Economics Letters*, *25*(1), 1–4. https://doi.org/10.1080/13504851.2017.1290763

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. *Conference on Innovations in Theoretical Computer Science*. https://doi.org/10.1145/2090236.2090255

EAEA - European Association for the Education of Adults. (2019). New Skills Agenda for Europe - European Association for the Education of Adults. https://eaea.org/our-work/influencing-policy/monitoring-policies/new-skills-agenda-for-europe/

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295x.109.3.573

Else-Quest, N. M., & Hyde, J. S. (2016). Intersectionality in Quantitative Psychological Research: Theoretical and epistemological issues. *Psychology of Women Quarterly*, *40*(2), 155–170. https://doi.org/10.1177/0361684316629797

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Fiske, S., & Taylor, S. (1991). *Social cognition*. Mcgraw-Hill Book Company. https://archive.org/details/socialcognition0002fisk/mode/2up

Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, *1*(6), 261–262. https://doi.org/10.1038/s42256-019-0055-y

Friemel, T. N. (2016). The digital divide has grown old: Determinants of a digital divide among seniors. *New Media & Society*, *18*(2), 313–331. https://doi.org/10.1177/1461444814538648

Ge, Y., Knittel, C. R., MacKenzie, D., & Zoepf, S. (2016). Racial and Gender Discrimination in Transportation Network Companies. *RePEc: Research Papers in Economics*. https://doi.org/10.3386/w22776

Gleason, J. M. (1999). An accurate, non-iterative approximation for studentized range quantiles. *Computational Statistics & Data Analysis*, *31*(2), 147–158. https://doi.org/10.1016/s0167-9473(99)00002-x

Google. (2017). re:Work - Reduce the influence of unconscious bias with these re:Work tools. Retrieved March 3, 2023, from https://rework.withgoogle.com/blog/fight-unconscious-bias-with-rework-tools/

Google Play. (n.d.). Gboard - the Google Keyboard - Apps on Google Play. Retrieved November 5, 2022, from https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin

Government of Japan. (2018). Game App Developer in Her 80s Opens ICT World for Fellow Seniors. Retrieved February 15, 2023, from https://www.japan.go.jp/tomodachi/2018/spring-summer2018/game_app_developer.html

Government of Japan. (2019). Restaurant of Mistaken Orders Brings Smiles. Retrieved February 15, 2023, from https://www.japan.go.jp/tomodachi/2019/winter2019/restaurant_of_mistaken_orders.html

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553–561. https://doi.org/10.1037/pspa0000016

Greenwald, A. G., & Krieger, L. H. (2006). Implicit Bias: Scientific Foundations. *California Law Review*, *94*(4), 945. https://doi.org/10.2307/20439056

Guegan, D., & Hassani, B. K. (2018). Regulatory learning: How to supervise machine learning models? An application to credit scoring. *The Journal of Finance and Data Science*, *4*(3), 157–171. https://doi.org/10.1016/j.jfds.2018.04.001

Guimaraes, A. A. R., & Tofighi, G. (2018). Detecting zones and threat on 3D body in security airports using deep learning machine. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.1189345

Hajian, S., Bonchi, F., & Castillo, C. F.-D. (2016). Algorithmic Bias. *Knowledge Discovery and Data Mining*, 2215–2216. https://doi.org/10.1145/2939672.2945386

Hao, K. (2019). This is how AI bias really happens — and why it's so hard to fix. Retrieved December 10, 2022, from https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Harnois, C. E. (2014). Are Perceptions of Discrimination Unidimensional, Oppositional, or Intersectional? Examining the Relationship among Perceived Racial–Ethnic-, Gender-, and Age-Based Discrimination. *Sociological Perspectives*, *57*(4), 470–487. https://doi.org/10.1177/0731121414543028

Harvard Sentences. (n.d.). Harvard Sentences. Retrieved December 5, 2022, from https://harvardsentences.com/#h5-harvard-sentences

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in machine learning - what is it good for? *European Conference on Artificial Intelligence*, 3–10. http://ceur-ws.org/Vol-2659/hellstrom.pdf

Hobson, S., & Dortch, A. (2022). IBM Policy Lab: Mitigating Bias in Artificial Intelligence. Retrieved February 9, 2023, from https://www.ibm.com/policy/mitigating-ai-bias/

Hultsch, D. F., Hertzog, C., Small, B. J., & Dixon, R. A. (1999). Use it or lose it: Engaged lifestyle as a buffer of cognitive decline in aging? *Psychology and Aging*, *14*(2), 245–263. https://doi.org/10.1037/0882-7974.14.2.245

Hunsaker, A., & Hargittai, E. (2018). A review of Internet use among older adults. *New Media & Society*, *20*(10), 3937–3954. https://doi.org/10.1177/1461444818787348

Keswani, V., Lease, M., & Kenthapadi, K. (2022). Designing Closed Human-in-the-loop Deferral Pipelines. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2202.04718

Kleinberg, J. (2018). Inherent Trade-Offs in Algorithmic Fairness. *Performance evaluation review*, *46*(1), 40. https://doi.org/10.1145/3292040.3219634

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, *10*, 113–174. https://doi.org/10.1093/jla/laz001

Lagacé, M., Charmarkeh, H., Laplante, J., & Tanguay, A. N. (2015). How Ageism Contributes to the Second-Level Digital Divide: The Case of Canadian Seniors. *Journal of technologies and human usability*, *11*(4), 1–13. https://doi.org/10.18848/2381-9227/cgp/v11i04/56439

Landers, R., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*, *78*(1), 36–49. https://doi.org/10.1037/amp0000972

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252–260. https://doi.org/10.1108/jices-06-2018-0056

Li, B., Sainath, T. N., Sim, K. C., Bacchiani, M., Weinstein, E., Nguyen, P., Chen, Z., Wu, Y., & Rao, K. (2018). Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model. *arXiv (Cornell University)*. https://doi.org/10.1109/icassp.2018.8461886

Loebach, J. L., Pisoni, D. B., & Svirsky, M. A. (2010). Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(1), 224–234. https://doi.org/10.1037/a0017609

Malisiewicz, T., & Efros, A. A. (2008). Recognition by association via learning per-exemplar distances. *Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2008.4587462

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Ministry of Economy, Trade and Industry. (2020). Design Policy Handbook 2020. Retrieved February 15, 2023, from https://www.meti.go.jp/english/press/2020/0420_003.html

Mitzner, T. L., Boron, J. B., Fausset, C. B., Adams, A., Charness, N., Czaja, S. J., Dijkstra, K., Fisk, A. D., Rogers, W. A., & Sharit, J. (2010). Older adults talk technology: Technology usage and attitudes. *Computers in Human Behavior*, *26*(6), 1710–1721. https://doi.org/10.1016/j.chb.2010.06.020

Nellis, S. (2021). Apple sees revenue growth accelerating after setting record for iPhone sales, China strength. Retrieved December 5, 2022, from https://www.reuters.com/article/us-apple-results/apple-tops-wall-street-expectations-on-record-iphone-revenue-china-sales-surge-idUSKBN29W2TD

Niehaves, B., & Plattfaut, R. (2014). Internet adoption by the elderly: employing IS technology acceptance theories for understanding the age-related digital divide. *European Journal of Information Systems*, *23*(6), 708–726. https://doi.org/10.1057/ejis.2013.19

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for t-tests through multivariate hypothesis. *L. K. Edwards (Ed.), Statistics: Textbooks and monographs, Applied analysis of variance in behavioral science*, *137*, 297–344.

Older Adult Technology Services (OATS). (2022). Digital Equity - OATS. Retrieved February 8, 2023, from https://oats.org/digital-equity/

Olkin, R., & Pledger, C. (2003). Can disability studies and psychology join hands? *American Psychologist*, *58*(4), 296–304. https://doi.org/10.1037/0003-066x.58.4.296

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books. https://books.google.de/books?id=NgEwCwAAQBAJ&printsec=frontcover#v=onepage&q&f=false

Open Speech Repository. (n.d.). Open Speech Repository. Retrieved December 6, 2022, from http://www.voiptroubleshooter.com/open_speech/british.html

OpenAI. (2022). OpenAI Charter. Retrieved December 5, 2022, from https://openai.com/charter/

Palmore, E. (2001). The Ageism Survey: First Findings. *Gerontologist*, *41*(5), 572–575. https://doi.org/10.1093/geront/41.5.572

Park, Y., Patwardhan, S. V., Visweswariah, K., & Gates, S. (2008). An empirical analysis of word error rate and keyword error rate. *Conference of the International Speech Communication Association*. https://doi.org/10.21437/interspeech.2008-537

R: The Studentized Range Distribution. (n.d.). http://stat.ethz.ch/R-manual/R-patched/library/stats/html/Tukey.html

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2212.04356

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some Effects of Training on the Perception of Synthetic Speech. *Human Factors*, *27*(4), 395–408. https://doi.org/10.1177/001872088502700404

Shah, D., Schwartz, H. A., & Hovy, D. (2020). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *arXiv (Cornell University)*. https://doi.org/10.18653/v1/2020.acl-main.468

Simon, J. (2022). Amazon SageMaker Clarify Detects Bias and Increases the Transparency of Machine Learning Models | Amazon Web Services. Retrieved February 9, 2023, from https://aws.amazon.com/blogs/aws/new-amazon-sagemaker-clarify-detects-bias-and-increases-the-transparency-of-machine-learning-models/

Smith, G., Pisoni, D. B., & Kronenberger, W. G. (2019). High-Variability Sentence Recognition in Long-Term Cochlear Implant Users. *Ear and Hearing*, *40*(5), 1149–1161. https://doi.org/10.1097/aud.0000000000000691

Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., & Mislove, A. (2018). Potential for Discrimination in Online Targeted Advertising. *HAL (Le Centre pour la Communication Scientifique Directe)*. https://hal.archives-ouvertes.fr/hal-01955343/file/Speicher-etal_PotentialAdDiscrimination_FAT2018.pdf

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. *Communications of The ACM*, *64*(8), 44–49. https://doi.org/10.1145/3464903

Storage, D. (2021). Stereotypes vs. Prejudice vs. Discrimination. https://www.youtube.com/watch?v=6Hr2XpBc_B4

Stypińska, J. (2021). Ageism in AI: new forms of age discrimination in the era of algorithms and artificial intelligence. *Proceedings of the 1st International Conference on AI for People: Towards Sustainable AI, CAIP 2021, 20-24 November 2021, Bologna, Italy*, 39. https://doi.org/10.4108/eai.20-11-2021.2314200

Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*, *2*(8).

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2011.5995347

Weiss, R. M., Bass, S. A., Heimovitz, H. K., & Oka, M. (2005). Japan's silver human resource centers and participant well-being. *Journal of Cross-Cultural Gerontology*, *20*(1), 47–66. https://doi.org/10.1007/s10823-005-3797-4

World Health Organization. (2022). Global report on ageism. (License: CC BY-NC-SA 3.0 IGO). Retrieved February 3, 2023, from https://www.who.int/teams/social-determinants-of-health/demographic-change-and-healthy-ageing/combatting-ageism/global-report-on-ageism

Xin, D., Ma, L., Liu, J., Macke, S., Song, S., & Parameswaran, A. (2018). Accelerating Human-in-the-loop Machine Learning. *International Conference on Management of Data*. https://doi.org/10.1145/3209889.3209897

Zafar, M., Valera, I., Rodriguez, M. J., & Gummadi, K. P. (2017). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *The Web Conference*, 1171–1180. https://pure.mpg.de/pubman/item/item_2422970_1/component/file_2422969/arXiv%3A1610.08452.pdf

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B. J., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J. A., & Perrault, C. R. (2021). The AI Index 2021 Annual Report. *arXiv (Cornell University)*. http://arxiv.org/pdf/2103.06312.pdf

Zhang, J., Wu, J., Qiu, Y., Song, A., Li, W., Li, X., & Liu, Y. (2023). Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review. *Computers in Biology and Medicine*, *153*, 106517.

Zhang, S. (2015). The "Harvard Sentences" Secretly Shaped the Development of Audio Tech. https://gizmodo.com/the-harvard-sentences-secretly-shaped-the-development-1689793568

Zhou, B., Yao, Y., & Luo, J. (2014). Cost-sensitive three-way email spam filtering. *Journal of Intelligent Information Systems*, *42*(1), 19–45. https://doi.org/10.1007/s10844-013-0254-7

Zickuhr, K., & Smith, A. (2012). Older adults and internet use. http://www.pewinternet.org/2012/04/05/older-adults-and-internet-use/

Zimmermann, M. (1986). *Neurophysiology of Sensory Systems*. Springer Nature. https://doi.org/10.1007/978-3-642-82598-9_3